An Examination of the Statistics and Reporting Procedures Used in the Standardized Testing Program of the Edmonton Catholic Schools

by

D.D. Harley & T.O. Maguire

Alberta Education

Alberta

89-02162/

An Examination of the Statistics and Reporting Procedures Used in the Standardized Testing Program of the Edmonton Catholic Schools

by

D.D. Harley & T.O. Maguire

This study was authorized by W.G. Hart
Edmonton Catholic Schools
Under contract to Alberta Education
Edmonton, Alberta

June 30, 1988

Please Note

The views and recommendations expressed in this report are those of the researchers and are not necessarily those of Alberta Education.

Acknowledgements

The Edmonton Roman Catholic Seperate School Board wishes to acknowledge with thanks, the work of the following persons who undertook and carried out this study for the school district.

Dr. Thomas O. Maguire,	Faculty of Education University of Alberta
Dr. Dwight D. Harley,	Faculty of Education University of Alberta

Abstract

The Edmonton Roman Catholic Separate School District received funds from Alberta Education to examine the statistical procedures which are used to produce records and reports from the district's standardized testing program. Reports and procedures were examined at three levels: student, school, and system. At the student level, reports on ability, achievement, and discrepancy assessment were evaluated. It was recommended that percentile ranges based on the standard error of measurement be used as the basic reporting statistic. For school and system level reporting, graphical procedures such as percentile plots and box and whisker displays were suggested as enhancements to current reports.The study concluded with a description of current and proposed statistical calculations.

If the recommendations of the report are followed, users of standardized test information will be constantly confronted with an appropriate level of uncertainty as they interpret student scores . Use of graphical displays should make it easier for readers to understand summary data thereby improving the quality of decisions emanating from the program.

Table of Contents

	List of Tables	vi
	List of Figures	vii
	Appendicies	viii
Chapter 1:	Background	1
Chapter 2:	Student Level Reporting	2
	Assessment	2
	Screening	4
Chapter 3:	System Level Reporting	8
	Residual Percentile Plot	10
	Box and Whisker Plot	11
Chapter 4:	Classroom or School Level Reporting	14
Chapter 5:	Analysis of Current and Proposed Statistical Procedures	16
	Standard Scores and the Standard Score Table	16
	Relationship Between Standard Scores and Percentile Ranks	18
	Calculation of Percentile Ranks	18
	Comparison of Methods of Obtaining Percentile Ranks	19
	Percentiles and Quartiles	20
	Summary Statistics Using Quartiles	21
	References	22

List of Tables

Table 1: Construction of Grade Equivalents. 3
Table 2: Standard Error for Currently Used Tests. 6
Table 3: Deciles of Appendix A Data Set. 8
Table 4: Five Number Summary of Appendix A Data Set. 11
Table 5: Five Number Summary of a Hypothetical Distribution. 12
Table 6: Comparison of Percentile Ranks. 19

List of Figures

Figure 1:	Percentile Comparison Plot.	9
Figure 2:	Residual Percentile Plot.	10
Figure 3:	Single Box and Whisker Plot.	12
Figure 4:	Multiple Box and Whisker Plot.	12
Figure 5:	Comparisons Between System, School(s) and Students.	15

Appendices

Appendix A: Sample Set of Results 23

Chapter 1
Background

The testing program for Edmonton Catholic Schools was recently reviewed by Mulcahy and others, and in their report indicated that standardized testing has a role at the "General Level" of school assessment. This level has four purposes: assessing strengths and weaknesses in areas such as reading and math, monitoring student progress, screening possible learning difficulties, and reporting to parents. Three of these purposes can be described as assessment (assessing strengths and weaknesses, monitoring progress and reporting to parents) because they involve the direct interpretation of test results. In the present context, screening is distinct from assessment since it involves the idea of discrepancy between two tests.

The standardized tests in current use by the system are norm-referenced. This means that the basis for their interpretation lies with how well a child does in comparison with his or her peers, or with students who participated in the norming study. In the various tests used by Edmonton Catholic Schools, four kinds of scores are produced (not all are reported for a single test). These are: raw scores (the total number of items answered correctly), grade equivalents, percentiles, and IQ or other scaled scores. The decision has been made on the purposes for the testing program and on the test that are used. The intent of the present paper is to examine reporting procedures and the statistics which contribute to them. Since there are three levels for reporting (student, class and system) and at the student level there are two purposes for testing (assessment and screening) it seems reasonable to examine the four kinds of reports separately. We believe that three characteristics should guide the examination. Reports should be simple so that they can be understood easily by all who use them, they should be useful so that they contribute to the educational goals and they should have integrity so that information contained in the report is a valid representation of the student's performance which leads to appropriate decisions and actions.

Chapter 2
Student Level Reporting

Assessment

At present, individual results are reported as IQ and percentiles for ability tests and grade equivalents and percentiles for achievement tests. It is our contention that a percentile range based on locally developed norms should be used as the principal means of conveying individual student information. There are three reasons for making this suggestion:

i. Indicating a student's ability or achievement level by means of a range or interval of scores properly conveys the uncertainty that is associated with the estimate. The reliability of a test indicates that a student's score is subject to many influences that tend to give a fuzzy picture of achievement or ability. Using the standard error of measurement, we can indicate the range of scores in which we would expect the student's true score to fall. For example if a student fell at the 35th percentile, we might report that we are 66% sure that the student's ability at this time lies somewhere between the 30th and 38th percentiles. This can be indicated on the student profile card using an interval rather than a bar graph.

ii. The percentile is chosen in preference to the grade equivalent score because we believe that grade equivalents have connotations or surplus meanings associated with them that are inappropriate.

To understand the problem with grade equivalent scores, it is instructive to examine the process used by the test developer in producing the grade equivalent scale. When a test is being normed, it is administered to a national sample of students. Suppose the test is administered in May. This is the ninth month of the school year. The average raw score for children in the third grade is assigned a grade equivalent of 3.9, the average raw score for children in the fourth grade is assigned a grade equivalent of 4.9, and the average raw score for children in the fifth grade is given a grade equivalent of 5.9. If the three grade means were 20, 27, and 35, then intermediate raw scores are assigned grade equivalents through interpolation on a ten month scale as shown below in Table 1.

Test items are written to maximize the discrimination among students. That means that the test score will spread the students out as much as possible. Thus, even though the average grade equivalent for fourth grade students in May is 4.9, the individual grade equivalent scores could be spread from 2 to 7. The grade equivalent score is a convenient statistical indicator of how well a student compares with the average student at that grade level, but it is not a good indicator of how well a student has mastered the important objectives of the curriculum, and it does not give benchmarks for determining adequate achievement. Unfortunately, the term, grade equivalent makes it seem like a benchmark interpretation is appropriate. For example, a teacher may notice that half the students are working below grade level. But this situation is as much a consequence of the way in which the test is constructed as it is an indication of teaching or learning problems. A fourth grade student who is tested in September (4.0) could easily have a grade equivalent of 3.5, and yet in terms academic development be progressing at an acceptable rate.

The score of 3.5 merely tells us that his performance level is below the average of fourth grade students in September.

Table 1: Construction of Grade Equivalents.

Mean Score	G.E.	Interpolation
20	3.9	
21		4.1
22		4.2
23		4.4
24		4.5
25		4.6
26		4.7
27	4.9	
28		5.0
29		5.1
30		5.2
31		5.3
32		5.4
33		5.6
34		5.7
35	5.9	

Since the grade equivalent is essentially a comparative scale with great potential for misinterpretation when applied to an individual student, it would seem more appropriate to use percentile ranks to convey the information. Percentile ranks give an accurate picture of relative standing without any of the confusing connotations. Indeed, handouts prepared by the The Edmonton Catholic School Board use the percentile to explain the notion of grade equivalent. It seems to us that locally derived percentile ranks where an adequate local norming population exists give the teacher an accurate picture of the information that is contained in the test score.

In the Mulcahy study, many teachers expressed a preference for reporting student results in terms of grade equivalents. Some may argue that it is the change in score over the year that is important not the absolute level. For them a growth of one grade equivalent or ten months is an indication of appropriate progress. Unfortunately, an increase in raw score of about 7 or 8 items is all that is necessary to produce a change of one grade equivalent in most of the subtests. Few teachers or students would be comfortable in accepting the validity of a claim based on so few items.

Student achievement at the individual level is not well monitored by standardized tests. System tests, classroom tests and even provincial tests provide a much better basis for making decisions about the adequacy of a student's progress. The best use of standardized tests is to identify students with potential learning problems so that they can be examined more closely.

iii. One of the goals of the testing program is to identify students who show a discrepancy between achievement and ability. At the present time this is being done

using a regression definition, but it could be done more directly. At each grade level covered by the testing program there are over 1000 students being assessed. With that many students it is possible to develop local norms that would have a great deal of stability from one year to the next. If all ability and achievement scores were converted to percentile ranks using the local cohort, then they would be a common basis for making comparisons between scores. A range of percentile ranks would be calculated for each student's performance on each test. A student who is a possible underachiever would be one whose ability range level was higher than the achievement range. For example, if the ability range was between the 85th and 92nd percentiles, and the achievement range was 42 to 48, then the teacher would be advised to look at the student more closely to see what factors might be causing this anomaly. Ability and achievement ranges that overlapped (e.g. 68 to 75 and 62 to 70) would indicate consistent performance.

Screening

Currently a regression system is used for screening individual achievement discrepancies. The first step in the process is to develop a least squares regression equation for predicting Grade Equivalent Scores on subtests of the CTBS from three predictor variables: chronological age, sex, and IQ. Grade discrepancy is then defined as the difference between the observed grade equivalent and the predicted grade equivalent (which is referred to as grade expectancy). So far as we can tell, the actual calculations that are carried out in practice are incorrect. Sex which appears in the regression equation as a variable is treated in the calculation of the grade expectancy score as a constant. This would artificially inflate (or deflate) the expectancy scores of either boys or girls (depending on how they were coded in the original regression analysis and the sign of the weight), so that there would be a tendency to identify too many underachievers in one sex. At some grade levels, and with some achievement variables, this error is trivial, but at other times it can be as high as three months. New weights are calculated at each administration time, and guidelines are provided to assist the teacher in interpretation.

The present system is cumbersome, expensive and incorrect. Because the procedure is done separately for each grade level, the use of age in the equation does not seem important. It appears that the administration does not like to use sex in the equation, so all that is left is intelligence. No accommodation is made for psychometric error, and as a result it is difficult to determine when a student should be flagged for special attention. As noted above, using locally produced percentile ranges, the teacher can quickly spot performances that do not overlap (not only between ability and achievement, but between different achievement tests as well). This procedure would be considerably cheaper than developing a multiple regression format each year, and it would define "under" and "over" achievement in terms that are directly understandable by the teacher. The concepts of over and under achievement as accepted by the system, are essentially comparisons between scaled scores on ability and scaled scores on achievement. But because they are expressed as grade equivalents, the discrepancy scores have an illusion of criterion referencing.

In reporting Student Scores it is recommended that:

1. Local percentiles for ability and achievement tests and subtests should be developed each year using the formulas in Chapter 4.

2. Student scores should be reported as percentile ranges. These could be recorded as intervals on the student profile card. It is recommended that a 66% confidence interval be used. (The score plus or minus one standard error). The lower limit of the range expressed in raw score units is found by subtracting one standard error of measurement from the student's observed raw score. The upper level is obtained by adding one standard error to the student's raw score. The standard error obtained from the test manuals for CTBS and CCAT scores are shown in Table 2. The lower and upper limits can then be transformed into percentiles using the locally developed percentiles.

3. Achievement discrepancies can be recognized by finding non-overlapping ranges in ability and achievement. CTBS profile can be examined for non-overlapping ranges, as well as CCAT-CTBS discrepancies. It is suggested that the CCAT Verbal range can be used to evaluate CTBS Reading and Vocabulary, and CCAT quantitative can be used for Math Concepts and Problem Solving.

Table 2: Standard Error for Currently Used Tests.
CTBS Standard Errors of Measurement

Form 3

Grade	Level	Subtest	Standard Error (Raw Score Units)
3	9	Vocabulary	2.2
		Reading	3.0
		Math Concept	2.2
		Problem Solving	2.1
4	10	Vocabulary	2.8
		Reading	3.6
		Math Concepts	2.7
		Problem Solving	2.3
5	11	Vocabulary	2.9
		Reading	3.7
		Math Concepts	3.1
		Problem Solving	2.4
6	12	Vocabulary	3.0
		Reading	3.7
		Math Concepts	3.1
		Problem Solving	2.4
7	13	Vocabulary	3.0
		Reading	4.2
		Math Concepts	3.0
		Problem Solving	2.5
8	14	Vocabulary	3.2
		Reading	3.9
		Math Concepts	3.1
		Problem Solving	2.5

Table 2 (cont.): Standard Error for Currently Used Tests.
CTBS Standard Error of Measurement

Form 5

Grade	Level	Subtest	Standard Error (Raw Score Units)
1	7	Vocabulary	2.3
		Reading	3.2
		Math Concept	2.2
		Problem Solving	2.0
2	8	Vocabulary	2.2
		Reading	3.1
		Math Concepts	4.0
		Problem Solving	3.1

CCAT Standard Error of Measurement

Grade	Level	Subtest	Standard Error (Raw Score Units)
3	A	Verbal	4.0
		Quantitative	3.1
		Non Verbal	3.4
6	D	Verbal	4.1
		Quantitative	3.1
		Non Verbal	3.5
9	F	Verbal	4.1
		Quantitative	3.0
		Non Verbal	3.6

Chapter 3
System Level Reporting

Several reports are prepared for the system as a whole. Most if not all are based on the performance of all students at a particular grade level. One display which is relatively complex to produce is the "Distribution Table (Bell Curve Report)." It is our belief that this table is not of sufficient utility to warrant production. Essentially the goal of the display is to provide the administrator with an "at glance" impression of the relation between system performance and national performance as described in the norms. This can be done much more easily by comparing system performance with national norms at various points along the achievement continuum.

While we are reluctant to encourage the use of grade equivalent scores at the student level because of the misinterpretation risk, at the system level both the risk and the consequences of misinterpretation are considerably reduced. There are many knowledgeable staff within the central administration who can insure the accuracy of inferences drawn from grade equivalent information. Consequently, we are recommending its use for comparing system performance to national norms by plotting the grade equivalents of fixed percentiles. Cleveland (1985) suggests that this is a simple yet elegant way of comparing distributions across their entire range.

This type of plot features percentiles from one distribution plotted against the corresponding percentiles of a second distribution with a reference line y= x overlayed for reference.

The following example compares local normed values based on the grade 8 reading data in Appendix A with the published normed values from the examiners' manual.

Table 3: Deciles of Appendix A Data Set.

Percentile Rank	Raw Score	Local G.E.	Normed G.E.	Differences
5	24.5	6.4	5.9	.5
10	28.0	7.0	6.6	.4
20	33.5	7.6	7.3	.3
30	37.5	8.0	7.8	.2
40	41.5	8.3	8.1	.2
50	44.5	8.6	8.5	.1
60	48.5	8.9	8.8	.1
70	51.5	9.1	9.2	-.1
80	55.5	9.5	9.6	-.1
90	61.0	10.0	10.1	-.1
95	64.5	10.4	10.6	-.2

The raw score column contains the scores corresponding to percentile ranks from the local distribution (e.g. 5% of the students in the Edmonton system have raw scores of 24.5 or less). The Local G.E. column contains the grade equivalent

score from the Examiners Manual (24.5 corresponds to a grade equivalent of 6.4 according to the publishers norms).

The Normed G.E. column contains the grade equivalent scores from the publisher's tables that corresponds to a percentile rank of 5 (e.g. for the national sample, 5% of the students tested at the appropriate time of the year had grade equivalent scores of 5.9 or less).

Thus students at the fifth percentile in Edmonton had higher G.E. scores than students at the fifth percentile in the national sample.

In summary, grade equivalent scores (raw scores and IQ scores could similarly be used) that correspond to the 5th, 10th, 20th, 30th, 40th, 50th, 60th, 70th, 80th, 90th, and 95th percentile ranks of the comparative group and scores that correspond to the same percentile ranks in the norming sample were obtained. Local normed values were plotted on the ordinate with the published norms plotted on the abscissa. The line of equal grade equivalents (y=x) was overlayed.

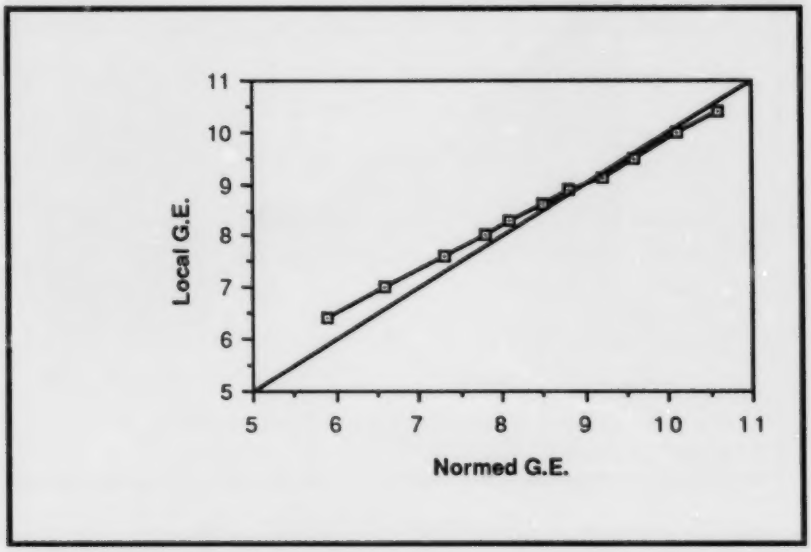

Figure 1: Percentile Comparison Plot.

The plot is interpretated by considering each data point in relation to the line y=x. Where the plot lies above the reference line, the local norms are superior the the published norms. Where the plot lies below the reference line the reverse is true.

From the plot we can see that the Edmonton Catholic students in the low and mid range area of reading achievement are above the national norms while those at the upper end of the scale are slightly below. The differences where they occur are less than half a year.

Residual Percentile Plots

Residual percentile plots are a slight variation of the previously described percentile comparison plots. They are used to summarize the same relationships and make the same comparisons. The two displays differ in regard to the data values that are plotted.

Once again consider the data presented in the preceding table. Rather than plotting pairs of norms, the differences or residuals between the local and the normed scores are calculated and are then plotted (ordinate) against the corresponding percentile ranks (abscissa).

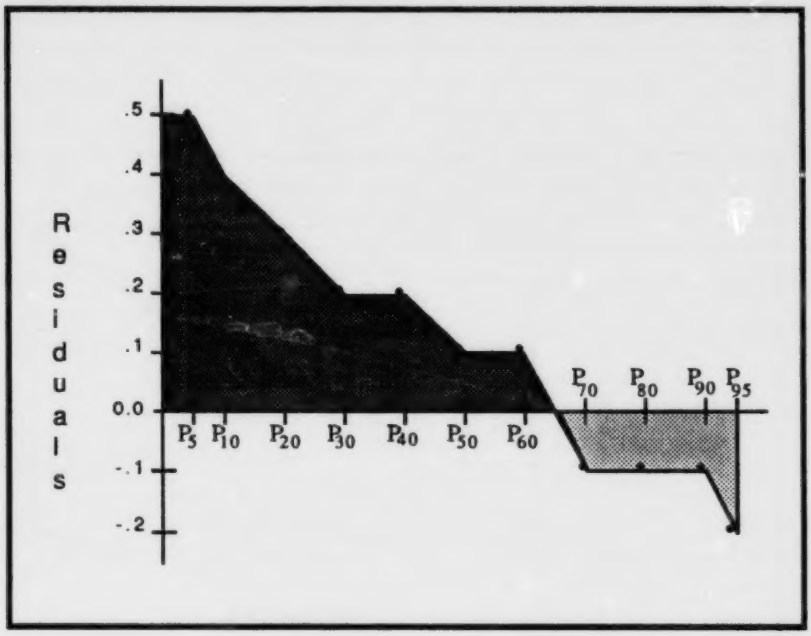

Figure 2: Residual Percentile Plot.

The interpretation of this style of plot is the same as that of the percentile comparison plot. If the data point is above the zero line then the local score is greater than the normed score. When the residual point is located in the lower half of the plot then the reverse is true. If the difference falls on the line y= 0 then there is no difference.

The previous two plots make the same comparison. The data are just viewed slightly differently. Percentile comparison plots provide a direct comparison of normed values of the two groups while residual plots emphasize the difference between the values. Differences can appear to be more dramatic when using residual plots however percentile comparison plots have a simplistic appeal. In any event there is no real advantage in choosing one type of comparative plot over the other. It is simply a matter of taste.

A second summary plot that would be useful for each grade would be box and whisker plots which compare the various scale scores to the national norms. A case can be made for using grade equivalents for these and previous plots since their purpose is to monitor performance against national values.

Both box and whisker plots and percentile difference plots can be used to compare system performance from one year to another. A detailed description of box and whisker plots is shown below.

Box and Whisker Plot

A practical method of comparing percentiles across sub-populations is by a graphical display known as a box and whisker plot. Tukey (1977) describes a box and whisker plot as a long thin box that stretches from the first to the third quartile, crossing it with a bar at the median. A line or whisker is drawn from each end of the box to the corresponding extreme.

Given a frequency and a cumulative frequency distribution, generation of a box-and-whisker plot is simple. Follow the procedure listed below.

1. Calculate the three quartiles. Recall that $Q_1 = P_{25}$, $Q_2 = P_{50}$ (or the median), and $Q_3 = P_{75}$.

2. Determine the minimum and maximum values of the distribution.

3. Plot a box (width is unimportant) with one end located at the first quartile (ie. Q_1) and the other located at the third quartile (ie. Q_3).

4. Cross the box with a line drawn at the median (ie. Q_2).

5. From each box end draw a line extending to the minimum or maximum value of the distribution. In some cases, as in the example below, the end point of the whisker indicating either the minimum or maximum value is crossed.

Figure 3 below is an example of a box and whisker plot built from the summary statistics of the sample data found in Appendix A.

Table 4: Five Number Summary of Appendix A Data Set.

$Q_1 = 35.10$ (first quartile)	Minimum value = 12.0
$Q_3 = 53.11$ (third quartile)	Maximum value = 75.0
$Q_2 = 44.69$ (median)	

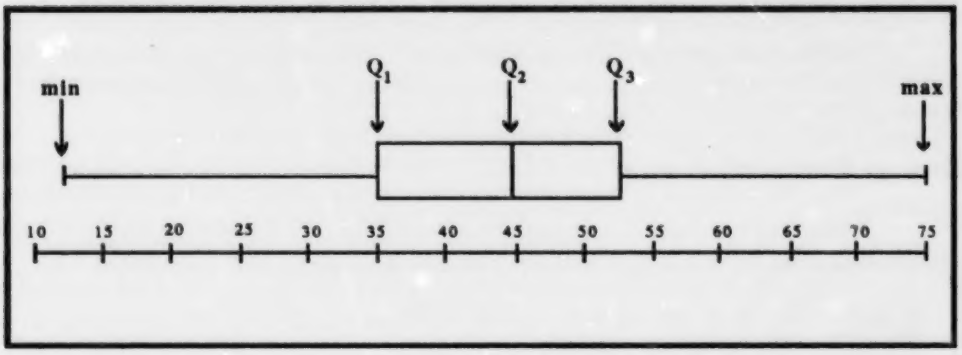

Figure 3: Single Box and Whisker Plot.

Box and whisker plots may be displayed either horizontally, as in the above example, or vertically. Several distributions can be plotted on a single set of axes in order to facilitate inter-distribution comparisons. Multiple box and whisker plots allow corresponding percentiles and symmetry to be compared visually across distributions.

An example of multiple box and whisker plots is presented below. Here the distribution used in the above example is compared with a distribution defined by some hypothetical values.

Table 5: Five Number Summary of a Hypothetical Distribution.

Q_1 = 40.00 (first quartile)	Minimum value = 20.0
Q_3 = 49.00 (third quartile)	Maximum value = 60.0
Q_2 = 47.00 (median)	

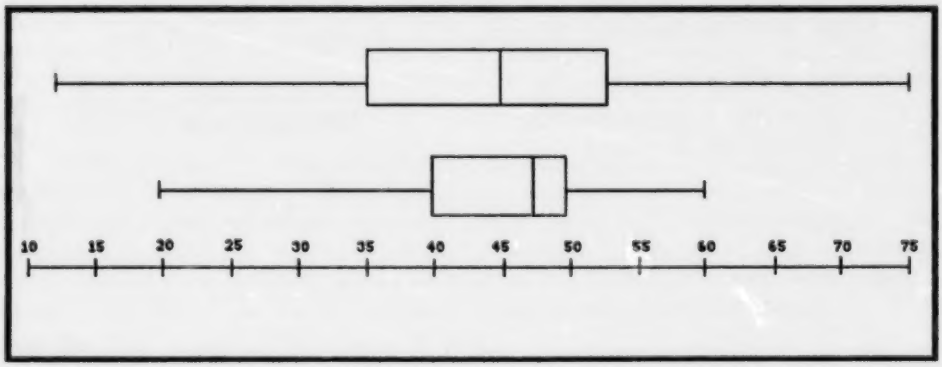

Figure 4: Multiple Box and Whisker Plot.

Generally speaking distributions are considered to be different if their boxes are non-overlapping. In the above example, the top distribution is more symmetric and the scores are more spread out than in the lower distribution. The median is slightly higher in the lower distribution. In spite of these differences in detail, the overall impression is that the two sets of data are quite similar.

Creating box and whisker plots by hand or with the aid of a micro computer is a trivial task but presents a considerable challenge on a mainframe computer. However with certain format modifications, the generation of box and whisker plots on mainframes simplifies. Procedures and algorithms for producing modified box and whisker plots on line printer can be found in Velleman and Hoaglin (1981).

It is recommended that percentile comparison plots, which show the relationship between system performance and national scores, be prepared for all CTBS subtests at all grade levels tested. Percentile comparison plots should also be developed for comparing system performance over time.

Chapter 4
Classroom or School Level Reporting

Teachers may be interested in how well their classes perform in relation to system and national norms. Box and whisker plots of classroom performance with system and national performance could be produced to allow the teachers to make this comparison. The three quartiles provide about as much definition as would be desirable in such a comparison. Again these could be carried out with grade equivalents since the inclusion of local and national values would force them to acknowledge the norm referenced basis of the data. It would even be possible to incorporate individual student ranges with the graph. An example is shown in figure 5 below.

It is recommended that summary statistics (mean, standard deviation) and comparative box and whisker plots be supplied to the schools. At the classroom level individual ranges could be incorporated into the display. Schools should also receive the reports listing individual student performance using ranges. Discrepancies between verbal performance and Reading-Vocabulary and quantitative performance and Math Concepts - Problem Solving could be flagged for attention.

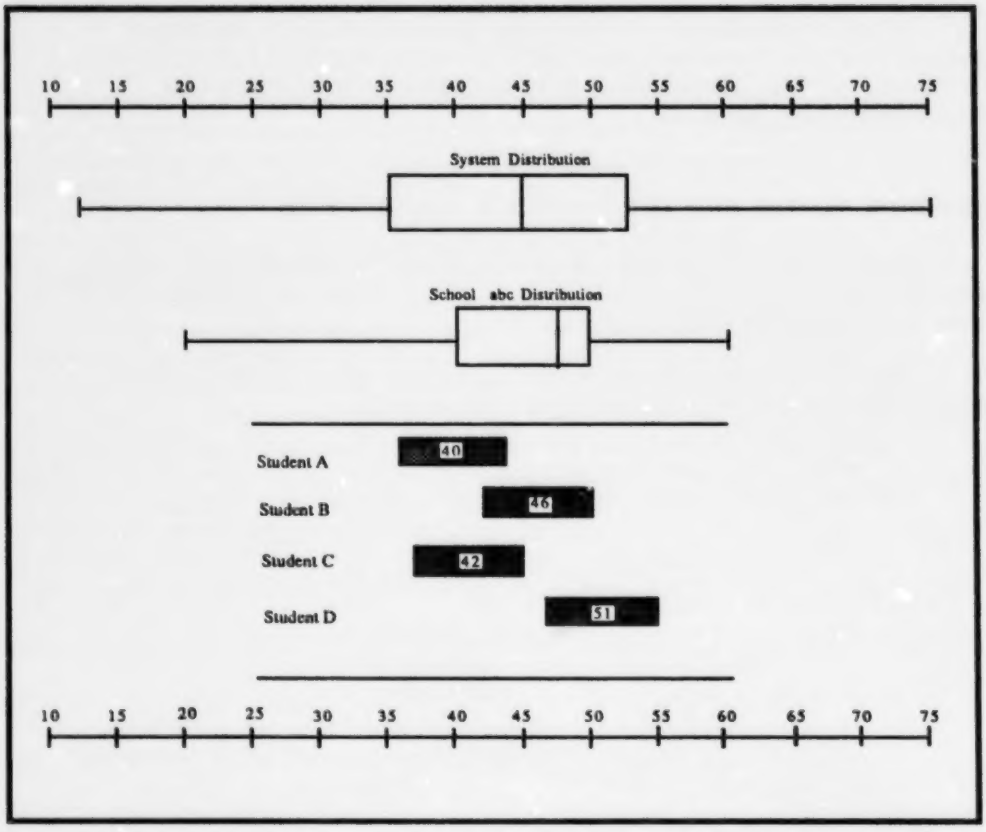

Figure 5: Comparisons Between System, School(s), and Students.

Chapter 5
Analysis of Current and Proposed Statistical Procedures

Several transformations and summary statistics are used in the current reporting program. These have been examined and are described and evaluated in this section. As well, computational procedures are outlined for proposed transformations and statistics.

Standard Scores and the Standard Score Table

Standard scores, commonly denoted as **Z**, are transformed scores and are assumed to be be interval in nature. Consequentlly, it can be appropriate to use **Z** scores in mathematical computations.

A linear (as opposed to normalized) standard score, Z_i, is produced by applying the following transformation to a raw score X_i.

$$Z_i=(X_i-m_x)/s_x$$

where

m_x is the mean of the distribution of N raw scores (ie. the first moment about the origin of the distribution), specifically

$$\mu_x=\frac{1}{N}\sum_{i=1}^{N}X_i$$

and

s_x is the population standard deviation of the distribution of N raw scores (ie. the second moment about the mean of the distribution), specifically

$$\sigma_x=\frac{1}{N}\left(\sum_{i=1}^{N}X_i-\mu_x\right)^2$$

Given that a frequency distribution of the N raw scores has been generated the following computational definitions for the mean and population standard deviation may be applied.

$$\mu_x=\frac{1}{N}\sum_{i=1}^{N}f_iX_i$$

where

f_i is the frequency corresponding to the raw score X_i.

$$\sigma_x^2 = \frac{\sum_{i=1}^{N} f_i X_i^2 - \left[\frac{1}{N}\left(\sum_{i=1}^{N} f_i X_i\right)^2\right]}{N}$$

where

$$\sigma_x = \sqrt{\sigma_x^2}$$

As standard scores are merely linear transformations of raw scores, their distribution has the same distribution shape as that of the raw scores. Therefore not all standard score distributions are normal distributions.

The algorithms for computing Z-scores were verified by examining output from Report Number IAS380-01 (Grade 8 Test Norm CTBS reading results for 1985/1986 - see Appendix A.). It is assumed that a lack of agreement between calculated values and tabled values found for the initial raw score values was caused by an imposition of an arbitrary minimum **Z** score value. The mean and standard deviation was calculated from the frequency distribution. The mean score matched that reported on the output, however the standard deviation appeared to be a sample standard deviation (ie. unbiased estimate with n-1 as the denominator) rather than the population standard deviation with n as the denominator (ie. biased estimate). As the data presented in the frequency represents a population the biased estimate should be used.

The Z-table consisted of the following elements:

Columns	Description
1 to 3	**Z** score
4 to 8	probability (ie. the area to the left) of the corresponding **Z**-score, P(X).
9 to 13	the ordinate(ie. the height) of the normal curve at the corresponding **Z** score, u(X).

The tabled probability entries were confirmed to four decimal places by comparing them to values computed from the rational approximation (Abramowitz & Stegun, 1972) given below:

$$P(X) = 1 - \frac{1}{2}\left(1 + \sum_{i=1}^{6} a_i X^i\right)^{-16}$$

where: $a_1 = 0.049867$ $a_2 = 0.021141$

$a_3 = 0.003278$ $a_4 = 0.000038$

$a_5 = 0.000049$ $a_6 = 0.000005$

for values of X such that $0 \leq X < \infty$

For values of X such that, $X<0$, $P(X)=1-P(|X|)$

The corresponding ordinate values (u(X))were similarly confirmed to four decimal places by evaluating the following expression for every tabled **Z** score value.

$$u(X) = \frac{1}{\sqrt{2\pi}} e^{-X^2/2}$$

for $0 \leq X \leq 4.00$

Relationship Between Standard Scores and Percentile Ranks

Percentile rank can be defined as the percentage of scores in a given distribution that falls below the midpoint of an interval containing a given score. It is a group specific statistic that is insensitive to general group performance. Because percentile ranks are rectangularly distributed, raw score differences near the middle of the distribution will be exaggerated as compared to the same raw score differences toward the extremes. As a result gains or losses of percentile ranks cannot be meaningfully compared for individuals at different points in the distribution. Percentile ranks are assumed to be an ordinal scale of measurement thereby precluding any mathematical calculation involving them. Contrary to a commonly held belief "it is usually considered inappropriate to base group comparisons on the average of individual percentile rank scores." (Crocker & Algina, 1986, pp. 442).

Upon inspection of the standardized testing procedures employed by the board, it is apparent that **Z**-scores are calculated for the sole purpose of obtaining percentile rankings. Procedures indicate that percentile ranks are taken to be the probability of a calculated standard score multiplied by 100 (ie. $Pr_i=P(Z_i)x100$).

This procedure does not generally result in percentile rank. What is actually being calculated is a percentile equivalent or a normalized percentile. The relationship between raw scores, **Z**-scores and percentile equivalents holds only in the case that the distribution of raw scores is normal. Interpretability and meaningfulness decrease as the distribution of raw scores departs from normality.

Calculation of Percentile Ranks

Crocker and Algina (1986, pp. 439) mathematically define percentile rank to be:

$$Pr_i = \frac{cf_{low} + .5(f_i)}{N} \times 100\%$$

where:

cf_{low} is the cumulative frequency of all scores less than the given score i.

f_i is the frequency of the given score i.

N the total number of people in the sample.

Pr_i is a proportion that corresponds to a given raw score, i, and hence ranges between 0.0 and 100.0.

The first step in calculating percentile ranks is to generate a frequency distribution of the raw scores. The second step is to generate a cumulative frequency distribution from the frequency distribution. Once these steps are complete it is just a matter of applying the above formula to determine the percentile rankings.

Comparison of Methods of Obtaining Percentile Ranks

To illustrate the different results consider the following example taken from Report Number IAS380-01 (Grade 8 Test Norm CTBS reading results for 1985/1986).

Table 6: Comparison of Percentile Ranks.

Raw Score	R.S. Freq	Z-Norm	%-ile	Cum Freq
17	1	-2.26	2	10
18	3	-2.18	2	13

Under the table look up procedure presently implemented by the board, the percentile rank for a raw score of 18 is 2. This is assuming that the raw score distribution is normal.

However, under the computational method detailed in a previous section a different result is obtained. From the following computation the percentile rank for a raw score of 18 is

$$Pr_{18} = \frac{10 + .5(3)}{1691} \times 100\% = 0.68$$

A percentile rank of 0.68 differs vastly from a percentile rank of 2.0. The difference is solely a result of the departure from symmetry of the raw score distribution. This particular example was chosen at an extreme of the raw score distribution, but similar differences can be found throughout the remainder of the distribution.

It is recommended that the table look up procedure for determining percentile ranks be abandoned in favour of the computational method. Should this recommendation be accepted, the process of **Z** score calculation becomes unnecessary.

Percentiles and Quartiles

Percentiles are a member of a family of descriptive statistics known as quantiles. In general terms quantiles are points on a distribution that divide it into groups of known and equal proportions. Percentiles divide the distribution into 100 intervals containing an equal number of data points. Deciles divide the distribution into 10 groups containing an equal number of data points and quartiles divide the distribution into four groups containing an equal number of data points.

As one might guess deciles and quartiles can be defined in terms of percentiles In particular the nine deciles are the 10^{th}, 20^{th}, 30^{th}, 40^{th}, 50^{th}, 60^{th}, 70^{th}, 80^{th}, and 90^{th} percentiles. The three quartiles can be thought of as the 25^{th}, 50^{th}, and 75^{th} percentiles, and the median, the point that separates the distribution into two groups with an equal number of data points, is the 50^{th} percentile.

Assuming the existence of a frequency and cumulative frequency distribution and a score-class interval of 1 the following formula (Glass and Stanley, 1970) may be applied to calculate percentiles, P_p.

$$P_p = L + \frac{p_n - cf_L}{f}$$

where:

p_n is the p^{th} proportion of the total number of scores (n) in the distribution

L is the lower limit of the score interval containing the p_n^{th} score

cf_L is the cumulative frequency up to the lower limit of the score interval containing the p_n^{th} score

f is the frequency of the interval containing the p_n^{th} score.

Values of P_p are usually considered to be integers and are obtained by rounding the calculated value to the nearest integer value. Where the value of P_p is calculated to be greater than or equal to 99.5, P_p is truncated to 99

P_p is a raw score that corresponds to a given proportion of the distribution, p.

As an example once more consider the data presented in Appendix A. To calculate the 80^{th} percentile (ie. P_{80}) of the distribution, the above formula is applied as follows:

$$P_{80} = 54.5 + \frac{(.80)(1691) - 1312}{48} = 55.35$$

Deciles, quartiles, median, and for that matter any other quantile that can be expressed in terms of percentiles can be calculated by applying the above formula.

Summary Statistics Using Quartiles

As noted previously statistical computations such as mean and standard deviation involving percentiles and percentile ranks are mathematically unsound and devoid of meaning. Due to the ordinal nature of these statistics, the median is the best choice of a measure of central tendency while the interquartile range is the most appropriate statistic to index variability.

The median is defined as the 50^{th} percentile in a group of scores. In other words the median is the score which splits the distribution into two equal portions. It is calculated using the formula on page 20 Considering the example data in Appendix A the median (**M**) is calculated to be **M** = 44.69.

The interquartile range (Q_I) is defined as one half the distance between the third and first quartile (Glass & Stanley, 1970). In mathematical terms:

$$Q_I = Q_3 - Q_1$$

The first quartile (Q1) is the the 25^{th} percentile (ie.point of a distribution below which 25% of the scores fall). The third quartile (Q3) is the 75th percentile (ie. point of a distribution above which 25% of the scores lie). Given that the first and third quartiles of the data in Appendix A are $Q_1 = 35.10$ and $Q_3 = 53.11$ respectively, the interquartile range is $Q_I = 53.11 - 35.10 = 18.01$.

The interquartile range is interpreted simply as an interval containing the middle 50% of the scores.

References

Abramowitz, M., and Stegun, I.A. (eds.). (1972). Handbook of Mathematical Functions, New York: Dover Publications .

Cleveland, W. S. (1985). The Elements of Graphing Data, Monterey, California: Wadsworth.

Crocker, L., and Algina, J. (1986). Introduction to Classical & Modern Test Theory, New York: Holt, Rinehart, and Winston.

Glass, G.V., and Stanley, J.C. (1970). Statistical Methods in Education and Psychology, Englewood Cliffs, New Jersey: Prentice-Hall.

Tukey, J. W (1977). Exploratory Data Analysis, Reading Massachusetts: Addison-Wesley Publishing Co.

Velleman, P. F., and Hoaglin, D. C. (1981). Applications, Basics, and Computing of Exploratory Data Analysis, Boston, Mass: Duxbury Press.

Appendix A:
Sample Set of Results

The following is a data set taken from Report Number IAS380-01 for Grade 8 1985/1986 C.T.B.S. reading results. The table contains the same data as presented in the report plus one extra column headed "Calculated Percentile Rank." Data in this column represent the percentile rank of each raw score in the distribution as calculated from the definition of percentile rank presented in Chapter 5.

The descriptive statistics summarize the distribution of raw scores.

n	= 1691
Mean	= 44.14
Standard deviation	=12.11
First quartile (Q_1)	= 35.10
Second quartile (Q_2)	= 44.69
Third quartile (Q_3)	= 53.11
Maximum	= 75
Minimum	= 12

Report Number IAS380-01 Grade 8 - CTBS Reading 1995/1996

Raw Score (X)	Frequency (f)	Z Scores (Z)	Percentile Equivalent Ranks	Calculated Percentile Ranks	Cumulative Frequency (cf)
0	0	-3.67	1	0	0
1	0	-3.58	1	0	0
2	0	-3.50	1	0	0
3	0	-3.42	1	0	0
4	0	-3.34	1	0	0
5	0	-3.25	1	0	0
6	0	-3.17	1	0	0
7	0	-3.09	1	0	0
8	0	-3.01	1	0	0
9	0	-2.92	1	0	0
10	0	-2.84	1	0	0
11	0	-2.76	1	0	0
12	1	-2.68	1	0	1
13	1	-2.59	1	0	2
14	0	-2.51	1	0	2
15	3	-2.43	1	0	5
16	4	-2.35	1	0	9
17	1	-2.26	2	1	10
18	3	-2.18	2	1	13
19	3	-2.10	2	1	16
20	8	-2.02	3	1	24
21	13	-1.93	3	2	37
22	11	-1.85	4	3	48
23	17	-1.77	4	3	65
24	20	-1.69	5	4	85
25	22	-1.60	6	6	107
26	24	-1.52	7	7	131
27	26	-1.44	8	9	157
28	31	-1.36	10	10	188
29	29	-1.27	11	12	217
30	27	-1.19	12	14	244
31	29	-1.11	14	15	273
32	38	-1.03	16	17	311
33	34	-0.94	18	19	345
34	50	-0.86	21	22	395
35	46	-0.78	23	25	441
36	39	-0.69	25	27	480
37	31	-0.61	28	29	511
38	39	-0.53	31	31	550
39	38	-0.45	33	34	588
40	50	-0.36	36	36	638
41	50	-0.28	39	39	688
42	57	-0.20	43	42	745
43	43	-0.12	47	45	788
44	48	-0.03	50	48	836
45	50	0.05	51	51	886
46	52	0.13	55	54	938
47	44	0.21	58	57	982
48	44	0.30	61	59	1026
49	60	0.38	64	62	1086
50	52	0.46	67	66	1138
51	56	0.54	70	69	1194
52	44	0.63	73	72	1238
53	50	0.71	75	75	1288
54	24	0.79	78	77	1312
55	48	0.87	81	79	1360
56	37	0.96	83	82	1397
57	39	1.04	85	84	1436
58	29	1.12	88	86	1465
59	28	1.20	88	87	1493
60	22	1.29	89	89	1515
61	29	1.37	91	90	1544
62	24	1.45	92	92	1568
63	22	1.53	93	93	1590
64	21	1.62	94	95	1611
65	20	1.70	95	96	1631
66	15	1.78	96	97	1646
67	12	1.87	96	98	1658
68	8	1.95	97	98	1666
69	7	2.03	97	99	1673
70	7	2.11	98	99	1680
71	5	2.20	98	99	1685
72	1	2.28	98	99	1686
73	2	2.36	98	99	1688
74	2	2.44	99	99	1690
75	1	2.53	99	99	1691
76	0	2.61	99		1691
77	0	2.69	99		1691
78	0	2.77	99		1691
79	0	2.86	99		1691
80	0	2.94	99		1691